Hierarchical aesthetic quality assessment using deep convolutional neural networks

Yueying Kao[a], Kaiqi Huang[a,*], Steve Maybank[b]

[a]*CRIPAC & NLPR, Institute of Automation, Chinese Academy of Sciences, University of Chinese Academy of Sciences, Beijing, China*
[b]*Department of Computer Science and Information Systems, Birkbeck College, University of London, London, UK*

Abstract

Aesthetic image analysis has attracted much attention in recent years. However, assessing the aesthetic quality and assigning an aesthetic score are challenging problems. In this paper, we propose a novel framework for assessing the aesthetic quality of images. Firstly, we divide the images into three categories: "scene", "object" and "texture". Each category has an associated convolutional neural network (CNN) which learns the aesthetic features for the category in question. The object CNN is trained using the whole images and a salient region in each image. The texture CNN is trained using small regions in the original images. Furthermore, an A&C CNN is developed to simultaneously assess the aesthetic quality and identify the category for overall images. For each CNN, classification and regression models are developed separately to predict aesthetic class (high or low) and to assign an aesthetic score. Experimental results on a recently published large-scale dataset show that the proposed method can outperform the state-of-the-art methods for each category.

Keywords: Aesthetic image analysis, convolutional neural networks, scene, object, texture

1. Introduction

Aesthetic image analysis has attracted increasing attention recently in the computer vision community [1, 2, 3]. Automated models for assessing aesthetic image quality are useful in many applications, e.g., image retrieval, photo management, photo enhancement, and photography [4, 5]. It is also interesting to investigate the high-level perception of visual aesthetics. In the last decade, some studies have shown that data-driven approaches [6, 7, 8, 9, 10, 11, 2, 12, 13, 14, 15] can be used to assess the aesthetic quality of images, although such assessments are difficult, even for humans. In early works, many

*Corresponding author.
Email addresses: yueying.kao@nlpr.ia.ac.cn (Yueying Kao), kqhuang@nlpr.ia.ac.cn (Kaiqi Huang), sjmaybank@dcs.bbk.ac.uk (Steve Maybank)

scene object texture

Figure 1: Example images of the category "scene", "object" and "texture".

handcrafted features were proposed based intuitions about how people perceive the aesthetic quality of images. These features include color [5, 6], the rule of thirds [6], content [7, 2], and composition [8]. Later, generic image descriptors such as Bag-Of-Visual-words (BOV) and Fisher Vectors (FV) were used to assess aesthetic quality. In [9] it is shown that the generic image descriptors can outperform the traditional handcrafted features. More recently, deep convolutional neural networks (CNNs) have been successfully applied to aesthetic quality assessment [16, 17]. CNNs can extract powerful features thus we use them in this paper to learn features for aesthetic quality assessment.

Most existing methods for assessing the aesthetic quality of images [6, 5, 7, 18, 8, 9, 12, 16, 17] treat all images equally without taking into account the diversity in image content or type. However, Oliva et al. [19] discriminate "scene" from "object" and "texture". They design a GIST descriptor for scene recognition. Considering that scene recognition, object recognition, and texture recognition are studied separately, the three categories should be treated differently for aesthetic quality assessment. In this paper, we classify all images into three categories, namely "scene", "object" and "texture". Figure 1 shows example images for the three categories, which suggest the different spatial layouts and fixated points in them. "Scene" images are composed of numerous objects, textures and colored regions, which are arranged in a variety of spatial layouts [20, 19]. All the elements in the scene may influence the humans' aesthetic judgments in ways which have been studied by psychologists [21, 22]. Object images generally contain a large salient object, which attracts the attention of a human viewer and may be a key factor for the assessment of visual aesthetics [7, 23]. Texture images have some statistical properties, and may contain repeating structures [24, 25, 26]. Humans may have different criteria for assessing the aesthetics of images in the three categories.

The adoption of different photographic styles for the three categories emphasizes their differences. For example, professional photographers often reduce the depth of field (DOF) to shoot single objects to create close-up photographs for the category "object", in which the foreground is clear and the background is blurred [1]. However, in photography for images in the category "scene", landscapes shot

with a narrow DOF are not considered pleasing; Instead, photographers prefer to have the foreground, middle ground, and background all in focus [1]. It is likely that the three categories may have different aesthetic criteria for human perception. Therefore in this paper different convolutional networks are proposed to learn the features required to make aesthetic judgements about images in the "scene", "object" and "texture" categories.

Aesthetic quality assessment can be formulated as a classification problem or a regression problem. It is known that aesthetic quality is a subjective attribute of images and there is a lack of precise definition. In most previous work on the aesthetic quality of images, the image datasets are obtained by online photo-sharing communities and rated by members of the community. The average score of user ratings is usually taken as a measure of the aesthetic quality of an image and it is also used to label the image. Typically, aesthetic quality assessment is reduced to a classification problem, by thresholding the average score to create a high quality class and a low quality class [6, 5, 4, 27, 7, 9, 8]. The images between the two classes are discarded. Only a few related works [6, 28, 17] use regression problem to calculate an aesthetic score. Visual aesthetic quality assessment should be formulated as a regression problem and the results compared with the ratings made by the human visual system [27]. In this work, a classification model and a regression model are both developed for each of the three categories "scene", "object" and "texture".

Based on the considerations mentioned above and on our previous work [17], we propose a novel framework for visual aesthetic quality assessment. Firstly, each image is assigned to one of the three categories "scene", "object" and "texture". Then, for each category, a specific convolutional neural network is constructed to learn aesthetic features automatically and to assess the aesthetic quality of an image. The aesthetic quality is described using a class (high or low) and a numerical score. In addition, a single CNN is also developed for the aesthetic quality assessment and the category recognition simultaneously for overall images. The CNN is simple and can also simultaneously consider the aesthetic labels and the different categories of images in contrast with the three specific CNNs. Experimental results on the recently published large-scale AVA dataset [29] demonstrate the effectiveness of our framework. Both of our classification and regression methods outperform the state-of-the-art methods for each category and our regression methods can achieve comparable results to our classifications.

The main contributions of our proposed method are summarized as follows.

- Inspired by the difference ways in which humans make aesthetic judgements and by the adoption of particular photographic techniques depending on the nature of the images, we propose a novel framework for visual aesthetic quality assessment by dividing images into three categories: "scene", "object" and "texture".

- Three specific CNNs, namely Scene CNN, Object CNN and Texture CNN, are constructed. The CNNs learn aesthetic features automatically. Moreover, a single CNN, namely A&C CNN, is also developed to learn effective features simultaneously for two targets: the aesthetic quality assessment and the category recognition.

- Each CNN classifies an image from the appropriate class according to its aesthetic level (high or low) and also uses regression to assign to the image a numerical score of its aesthetic quality.

The rest of this paper is organized as follows. In Sec. 2, the related works are summarized. The methods for aesthetic quality assessment are described in detail in Sec. 3. Sec. 4 describes the experimental setup and results. Finally we conclude the paper in Sec. 5.

2. Related work

Most previous works [6, 5, 8, 9, 18] on aesthetic image analysis focus on the challenging problem of designing appropriate features. Typically, handcrafted features are proposed based on intuitions about human perception of the aesthetic quality of images. For example, Datta et al. [6] design certain visual features such as colorfulness, the rule of thirds, and low depth of field indicators, to discriminate between aesthetically pleasing and displeasing images. Dhar et al. [8] extract some high level attributes including compositional, content, and sky-illumination attributes, which are characteristically used by humans to describe images. In [9] generic image descriptors such as BOV or FV are used to assess aesthetic quality. It is shown that they can outperform the traditional handcrafted features. More recently, deep convolutional neural networks have been successfully applied to many visual tasks. CNNs learn powerful features automatically. For instance, in [30, 31, 32] it is demonstrated that CNNs achieve the state-of-the-art results in visual classification task on ImageNet. CNNs have been applied to aesthetic quality assessment [16, 17]. The CNNs learn features for the automatic aesthetic classification of all images and obtain the state-of-the-art performance. There are some key differences between the work in [16, 17] and our work. Firstly, all images are treated equally in [16, 17], while we divide the images into three categories. Secondly, they design CNNs for all images without considering their types, whereas we train specific networks with different architecture and inputs for each of the three categories. We also train a single CNN with the supervision of aesthetic and category labels for overall images. Thirdly, the objective of their work is to assess the aesthetic class: high or low. In contrast, we obtain a numerical score for the aesthetic quality of an image.

All of the related works discussed above do not consider different types of images. In contrast, Luo et al. [18] and Tang et al. [2] propose a content-based photo quality assessment method in which

the images are divided into seven categories ("animal", "plant", "static", "architecture", "landscape", "human" and "night") based on their content. New visual features are developed for that different categories. They consider that professional photographers may adopt different photographic techniques and may have different aesthetic criteria in mind for each type of image. The human perception of the aesthetics of visual textures is studied experimentally in [33]. Thumfart et al. [33] model the relationship between computational texture features and aesthetic properties of visual textures. In our work we divide the images into three categories: "scene", "object" and "texture" based on the composition, layout and photographic styles of images. Different CNNs are developed to assess the aesthetic quality for each category. For the category "scene", a global view is used to train the CNNs. The global view and the saliency region are used to train the CNNs for the category "object". Local regions are used to train the CNNs for "texture".

It is known that aesthetic quality assessment can be formulated either as a classification problem or as a regression problem. In most previous works classification models are adopted to predict "high quality" or "low quality" classes [6, 5, 4, 27, 7, 9, 8]. However, the classification model does not provide a numerical score of the aesthetics of an image. Thus visual aesthetic quality assessment should be a regression problem similar to its rating process in human visual system [27]. Some related literatures [34, 35, 36] have show that methods based on human visual system can improve the quality of images effectively. Datta et al. [6] and Wu et al. [28] assess aesthetic quality using a regression model and obtain some success. However, their regression model cannot obtain comparable results to existing works on classification. In contrast, our previous regression model [17] outperforms the existing methods for classification. In our work, both the aesthetic class and the score are predicted using both classification and regression models. The classifier outperforms previous state-of-the-art classifiers. Besides, the regression results can obtain comparable results to our classification.

3. Our approach

The proposed framework is illustrated in Fig. 2. Firstly, we divide all images into three categories: "scene", "object" and "texture". Then, for each category, a specific convolutional neural network is trained to learn aesthetic features automatically, shown in Fig. 2(a). Furthermore, the framework can also be proposed with a single CNN (shown in Fig. 2(b)). The CNN can simultaneously assess the aesthetic quality and identify the category for overall images. The classification and regression models are incorporated into the networks separately. We explain the aesthetic features learned by the specific networks for each category and overall images.

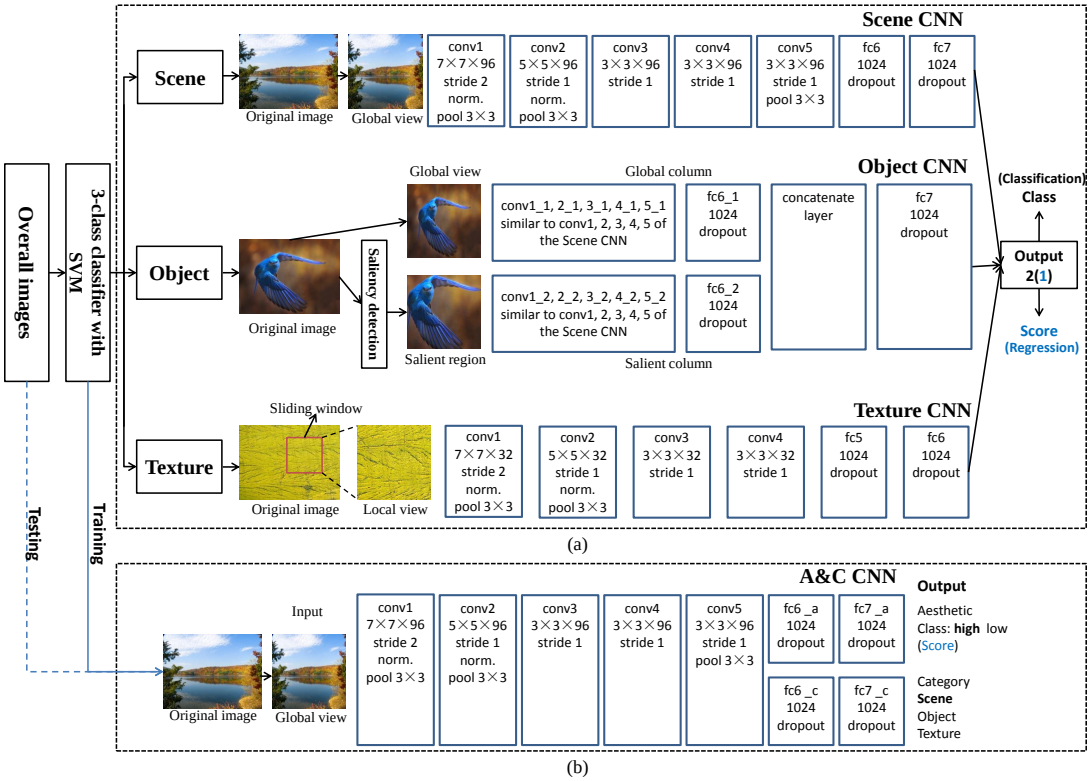


Figure 2: Framework for measuring image aesthetics. (a) The framework with three specific CNNs for different categories. (b) The framework with a single CNN for overall images.

The images can be classified as "scene", "object" or "texture" manually or automatically. Because the dataset used in this paper is very large scale, giving all the images category labels manually is expensive and time consuming. Thus we train a 3-class linear SVM classifier to divide all the images automatically by labeling part of the dataset and extract the features of layer fc7 (CNN-fc7) by utilizing the network pre-trained on ImageNet classification task [30]. The CNN-fc7 features are computed by passing the mean-subtracted images through five convolutional layers and two fully-connected layers with forward propagation. More details of CNN-fc7 are in [30]. The CNN-fc7 features have great representation power and have been applied to many tasks, such as image classification [30] and object detection [37].

3.1. Our framework with three specific CNNs

For the aesthetic quality assessment of the three categories, a simple idea is that three specific convolutional neural networks are proposed. 1) Scene CNN: for "scene", a single-column CNN is trained on the inputs with global view, which represents an entire image. 2) Object CNN: a two-column

Original image Warped image Padded image

Figure 3: Global views of an example image and of warped and padded versions of the image.

CNN is proposed with two inputs, global view and salient region for "object". 3) Texture CNN: for "texture", a single-column CNN is presented with a patch from the image. Classification and regression models are trained separately for each CNN to assess the aesthetic quality of the images in each class.

3.1.1. Scene CNN

Many important factors affect human assessments of the aesthetics of images. These factors include color [38, 39, 40], composition [22], visual attention [41], familiarity [6] and visual complexity [42]. Most of these factors have been used to design features suitable for the assessment of image aesthetics. However, they do not cover all possibilities. Convolutional networks can learn powerful features automatically. The features are much better than designed features for tasks such as visual classification [30] and aesthetic quality classification [16, 17]. Why not use convolutional networks to learn the features for assessing the aesthetics of "scene" images?

Figure 1 shows an example image in the category "scene". As every component in "scene" images plays a big role in the assessment of aesthetic quality, the global view of images is applied to train the network. The global view of an image is the entire image. For the implementation of our network, we normalize image sizes with two different transformations mentioned in [16], namely warping and padding, in order to obtain square images. Figure 3 shows an original image, a warped image and a padded image. The warped and padded images reflect the global view of original image.

The architecture of the Scene CNN is illustrated in Fig. 2(a). The network contains five convolutional layers and three fully-connected layers. Both the first and second convolutional layers are followed by max-pooling layers and response-normalization layers. The first convolutional layer filters the $227 \times 227 \times 3$ input patch (extracted from the resized image $256 \times 256 \times 3$ randomly) with 96 kernels of size $7 \times 7 \times 3$ with a stride of 2 pixels. The second convolutional layer takes the response-normalized and pooled output of the first convolutional layer as the input and filters it with 256 kernels

of size $5 \times 5 \times 96$. Each of the third, fourth and fifth convolutional layers has 96 kernels of size $3 \times 3 \times 96$. The pooled output of the fifth convolutional layer is input to the first fully-connected layer. Each of the first and second fully-connected layers have 1024 nodes. The number of nodes in the third fully-connected layer and the objective function depend on the task of the CNN, which in this case is the assessment of image aesthetics.

A. Classification model: The classification model is implemented by the Scene CNN, by having only two nodes in the last fully-connected layer. Each image is labeled "high quality" or "low quality". As in [30], the softmax loss layer is adopted in the training phrase of the classification network. The output of the last fully-connected layer is fed to a 2-way softmax which produces a distribution over the 2 class labels. The objective of our network is to maximize the multinomial logistic regression, which is equivalent to maximizing the average across training cases of the log-probability of the correct label under the prediction distribution.

B. Regression model: The aesthetic quality assessment is also interpreted as a regression problem. There are two major reasons. Firstly, the regression model is a direct emulation of humans in the photo rating process and closer to the visual aesthetic quality assessment in the human visual system [27]. Secondly, the features learned by the convolutional networks may contribute to making the regression problem more solvable.

Our regression model is trained by the Scene CNN. The regression network contains five convolutional layers and three fully-connected layers. The average score of user ratings for aesthetic quality of each image is made the label of the image. The last fully-connected layer is set one node and a sum-of-squares layer is utilized as the loss function. The output of the last fully-connected layer corresponds to the predicted aesthetic score $\hat{y}$. Then, the output of the last fully-connected layer and the label of images are taken as inputs for the sum-of-squares layer with Euclidean space. The objective of this layer is to minimize the squared L2 norm of the difference between its inputs:

$$\min \sum_{i=1}^{n} ||y_i - \hat{y}_i||_2^2, \tag{1}$$

where y_i is the ground truth for image i, $\hat{y}_i$ is the predicted value of image i and the n is the number of image.

We initialize the weights and biases in all the convolutional layers and the first two fully-connected layers in the regression network with the parameters of the classification network. The same operation is also applied to the other two categories: "object" and "texture". The regression model then automatically predicts the aesthetic scores for test images.

3.1.2. Object CNN

"Object" images generally contain a large salient object, which is likely to attract attention and may play a big role in the assessment of visual aesthetic quality [7, 41, 23]. Furthermore, photographers tend to adopt particular photographic techniques (e.g., macro, shallow DOF) to focus on the object and make the image pleasing. Here we attempt to use the salient region and global view to assess the aesthetic quality of object images. The influence of the salient region on aesthetics of the image can be studied.

Firstly, the salient objects are detected via graph-based manifold ranking [43], and a bounding box is generated for the salient region. Figure 2(a) shows a salient region. Then the Object CNN with two columns automatically learns the aesthetic features and assess the aesthetic quality with two inputs, the global view and the salient region. The architecture of the Object CNN is illustrated in Fig. 2(a). The inputs of the network are global view and salient view, the convolutional layers and the first fully-connected layer of different columns are trained independently. Here f_g and f_s indicate two vectors which are taken from the first fully-connected layers of the two columns separately. We concatenate the two vectors to one vector $f = [f_g, f_s]$ and train the last two fully-connected layers jointly. The architectures of the two columns are the same, but their weights are not shared, because the two columns learn the global features and salient features separately. The last fully-connected layer with two nodes and softmax loss layer are used for classification, and the last fully-connected layer with one node and a sum-of-squares layer is used for the regression task.

3.1.3. Texture CNN

Experiments on humans aesthetic perception of visual textures are reported in [33]. Supervised machine-learning methods are used to model the relationship between computational texture features and aesthetic properties of visual textures. However, the texture features are designed manually. In this paper, we utilize the convolutional neural networks to learn the features.

An image can be considered as a visual texture when (1) there is significant variation in the intensity levels of nearby pixels [26] and (2) the image is stationary (i.e., under a proper window size, observable subimages appear similar) [25, 33]. The structure of "texture" images is different from the structure of "scene" and "object" images. The original texture image is represented by 16 patches of size 256×256. These patches are extracted from the image with a sliding window and is called local view here. An example of a patch (local view) is shown in Fig. 2. The different patches are added to the dataset.

The architecture of the single-column Texture CNN for aesthetic quality assessment of "texture" is illustrated in Fig. 2(a). The network contains four convolutional layers and three fully-connected layers.

Both of the first and second convolutional layers are followed by max-pooling layers and response-normalization layers. The first convolutional layer filters the $227 \times 227 \times 3$ input patch (extracted from a local view $256 \times 256 \times 3$ randomly) with 32 kernels of size $7 \times 7 \times 3$ with a stride of 2 pixels. The second convolutional layer takes (response-normalized and pooled) output of the first convolutional layer as the input and filters it with 32 kernels of size $5 \times 5 \times 32$. Each of the third and fourth convolutional layers has 32 kernels of size $3 \times 3 \times 32$. Each of the first and second fully-connected layers have 1024 nodes. The number of nodes in the third fully-connected layer and objective function are similar to the single column network for "scene".

For a test image, we average the probabilities of 16 local views and select the class with highest value for class prediction, and use the averaged score of 16 local views for aesthetic score prediction.

3.2. A&C CNN

For a test image, the framework with three specific CNNs assesses its aesthetic quality based on its category. Its category is identified with a SVM classifier. To remove the effect of the SVM classifier and reduce the parameters of the framework for practical application, we propose a simple framework (shown in Fig. 2(b)) with a single A&C CNN to simultaneously assess the aesthetic quality and recognize the category for overall images. The CNN can consider the aesthetic labels and the different categories of images in contrast with the three specific CNNs.

The architecture of the A&C CNN for aesthetic quality assessment of overall images is illustrated in Fig. 2(b). The input of the network is the global view similar to the Scene CNN. Since the network has two targets: aesthetic quality assessment and category identification, the network contains five convolutional layers for common features learning. With the supervision of the two labels, the features may be more effective for one or two tasks. Three fully-connected layers learn parameters separately and independently for each target. The setup of five convolutional layers and two fully-connected layers are similar to the Scene CNN. The number of nodes in the third fully-connected layer and objective function for aesthetic quality assessment task are similar to the single column network for "scene". For category identification task, the third fully-connected layer is fixed to three nodes and objective function is softmax loss function. For training, each image is labeled with aesthetic label and category label. For a test image, its aesthetic and category labels are predicted with this CNN.

4. Experimental results

In this section, we evaluate the proposed framework and other state-of-the-art methods on the recently published AVA dataset [29]. The experimental results show that our networks for each category

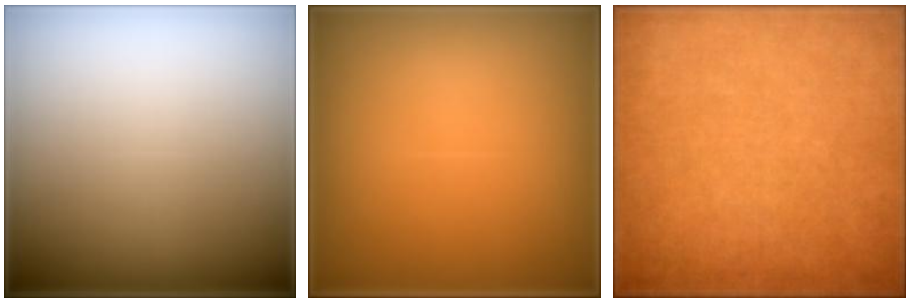


Figure 4: Mean images of the category "scene", "object" and "texture". The mean image is created by normalizing all the images of each category to the same size and averaging them.

can automatically assess the aesthetic quality (class and score) of images and can outperform existing state-of-the-art methods.

4.1. Dataset and evaluation

The AVA dataset [29] contains more than 255,000 images, each of which has about 200 voters for assessing the aesthetic score from one to ten. To divide the dataset into three categories: "scene", "object" and "texture", we manually labeled 5,000 images. Then we extract the CNN-fc7 features [30] on the AVA dataset. For comparison, GIST [19] descriptor, and Fisher Vector [44, 45] descriptor encoded from SIFT information (FV-SIFT) are also extracted. 3-class linear SVM classifiers are trained for each feature and evaluated using 5-fold cross-validation. The accuracy of these features on the labeled 5,000 images is shown in Table 1. The rest images are classified by the trained SVM classifier with CNN-fc7 features. In the AVA dataset 94,290 images are labeled with "scene" category, 155,612 images with "object", and 5,233 images with "texture". All the images of each category are used to generate the mean images shown in Fig. 4. The three mean images show the different spatial layout and fixated point in the three categories. To assess aesthetic quality on each category, 7,000 images in the category "scene", 13,000 images in "object" and 1,000 images in "texture" are selected randomly for testing, the rest in each category are for training.

For the classification task, the training set is divided into two classes: high quality images and low quality images, as in [29]. We designate the images with an average score larger than $5 + \delta$ as high quality images, those with an average score smaller than $5 - \delta$ as low quality images. Images with an average score between $5 + \delta$ and $5 - \delta$ are discarded. In order to make the classification problem easier we set the δ to 1 for the training set and set δ to 0 for the test set to obtain the ground truth labels for the three CNNs. Both $\delta = 0$ and $\delta = 1$ for the training set, and δ to 0 for the test set is fixed to obtain

Method	GIST	FV-SIFT	CNN-fc7
Accuracy (%)	76.0	83.8	91.3

Table 1: The accuracy of the SVM classifiers with different features on the manual labeled images for dividing images into three categories: "scene", "object" and "texture".

the ground truth labels for the A&C CNN. For the regression task, each image is assigned a label equal to the average score for aesthetics.

For evaluation, we compare the results of each specific CNN with the results obtained by other methods. Marchesotti et al. [9] show that generic image descriptors outperform the traditional descriptors. Therefore we implement the generic image descriptors in [9]: We extract GIST [9, 19] descriptor, Bag-Of-Visual-words [9, 46, 47] descriptor encoded from SIFT [9, 48] information (BOV-SIFT), and Fisher Vector [9, 44, 45] descriptor encoded from SIFT information (FV-SIFT) for each category. Some SVM classifiers and regression models are trained by LibSVM [49] using the extracted features. Furthermore, We implement the SCNN in [16] (here called Rapid SCNN) for each category. For comparison on regression task, the mean residual sum of squares error (MRSSE) [17] and Pearson's r correlation coefficient are applied to evaluate the results. In ideal conditions, the predicted values are equal to the ground truth, and Pearson's correlation coefficient $r = 1$. In the end, we also compare our framework with the existing methods on overall images (the whole dataset without considering their category label).

4.2. "Scene" results

To select a architecture for the category "scene", we train 10 single-column networks with different architectures and inputs. The difference mainly focus on the filter size of first convolutional layer and channels and number of all convolutional layers. Table 1 shows the architecture and performance of the 10 networks on the visual aesthetic quality classification task in detail. The results show that the architecture 6 # with the input of warped images performs best and the architecture 9 # with the input of padded images performs well. Thus we fix the Scene CNN with five convolutional layers and 3 fully-connected layers similar to the architectures 6 # and 9 #.

For evaluation on the classification task for "scene", we compare the results of Scene CNN and the A&C CNN with other methods, which are shown in Table 3. Here the A&C CNN is only evaluated on test "scene" images. We can see that the A&C CNN with input of warped images achieves the best performance, and both of our Scene CNNs with input of warped or padded images significantly outperform other methods. This suggests that our networks can learn relevant features and predict the aesthetic class effectively. Table 2 and 3 also reveal that the results obtained from warped images are

Arch No.	Input	conv1 (pool, rnorm) (filter size, channels)	conv2 (pool, rnorm) (channels)	conv3 (channels)	conv4 (channels)	conv5 (pool) (channels)	fc6 (nodes)	fc7 (nodes)	fc8 (nodes)	Accuracy (%)
1 #	warped	$11 \times 11, 48$	48	48	48	48	1024	1024	2	75.05
2 #	warped	$7 \times 7, 48$	48	48	48	48	1024	1024	2	75.38
3 #	warped	$7 \times 7, 64$	64	64	64	–	1024	1024	2	75.38
4 #	warped	$7 \times 7, 64$	64	64	64	64	1024	1024	2	75.77
5 #	warped	$7 \times 7, 96$	96	96	96	–	1024	1024	2	75.67
6 #	warped	$7 \times 7, 96$	96	96	96	96	1024	1024	2	**75.91**
7 #	padded	$7 \times 7, 48$	48	48	48	48	1024	1024	2	75.57
8 #	padded	$7 \times 7, 64$	64	64	64	64	1024	1024	2	75.37
9 #	padded	$7 \times 7, 96$	96	96	96	96	1024	1024	2	75.40
10 #	padded	$11 \times 11, 96$	192	384	256	256	1000	256	2	74.57

Table 2: Accuracy of 10 networks with different architectures and inputs for "scene".

better than those obtained from padded images.

For the regression task, we select the warped images as input. The classification model with the Scene CNN is utilized to fine tune the regression network. The A&C CNN with regression model is also only tested on the "scene" images. The MRSSE and Pearson's r correlation coefficient for each regression method for the class "scene" are shown in Table 4. The variance σ^2 of the ground truth in the "scene" test set is 0.5339. Datta et al. [6] demonstrate that the independent variables explain something about y if $MRSSE \leq \sigma^2$. Both of the Scene CNN and the A&C CNN on test "scene" images achieve $MRSSE < \sigma^2$, which shows that our methods are able to predict aesthetic scores. Moreover, we can see in Table 4 that the A&C CNN achieves the best performance and the Scene CNN obtains comparable results with the A&C CNN as measured by MRSSE and Pearson's r metric. Our networks predict the aesthetic scores automatically and effectively.

4.3. "Object" results

We make each column of the Object CNN similar to the Scene CNN. For evaluation on the classification task for "object", we compare the results of our networks with SVM classifiers using GIST, BOV-SIFT and FV-SIFT features, our implemented Rapid SCNN [16], and single-column networks with the input of warped images, padded images and salient region respectively. The single-column network has the same architecture as the Scene CNN. The results for "object" are shown in Table 3. It is apparent that the network using the architecture of the Scene CNN with input of warped images performs better than that with padded images. Table 3 also shows that the Object CNN with global view (here we select the warped images) and salient view obtains the best performance, the single-column CNN with global view obtains comparable results with the Object CNN, and the network with the ar-

Category	Method	Accuracy(%)
Scene	GIST	73.41
	BOV-SIFT	73.53
	FV-SIFT	73.60
	Rapid SCNN [16] (warped)	75.44
	Scene CNN (warped)	75.91
	Scene CNN (padded)	75.40
	A&C CNN (warped)	**76.04**
Object	GIST	68.78
	BOV-SIFT	68.67
	FV-SIFT	69.31
	Rapid SCNN [16] (global, warped)	72.84
	Single-column CNN (global, warped)	73.50
	Single-column CNN (global, padded)	72.92
	Single-column CNN (saliency)	71.47
	Object CNN (warped + saliency)	**73.66**
	A&C CNN (warped)	73.30
Texture	GIST	63.5
	BOV-SIFT	65.4
	FV-SIFT	63.5
	Rapid SCNN [16] (global, warped)	64.0
	CNN similar to Scene (patch)	67.9
	Texture CNN	68.8
	A&C CNN (warped)	**71.6**

Table 3: Performance comparison on aesthetic quality classification task for "scene", "object" and "texture" respectively.

chitecture of the Scene CNN using the salient region only can still significantly outperform the method with the generic image descriptors. This suggests that the salient region plays a big role in aesthetic assessment in the category "object" and the single-column network can also learn the features of salient region.

For the regression task, we utilize the classification model with the Object CNN to fine tune the regression network. The A&C CNN with regression model is also only tested on the "object" images. The MRSSE and Pearson's r correlation coefficient for each regression method for "object" are shown in Table 4. The variance σ^2 of the ground truth in the "object" test set is 0.5444. We achieve $MRSSE = 0.4092$ with the Object CNN, and $MRSSE = 0.3988$ with the A&C CNN. Moreover, Table 4 shows that our A&C CNN achieves the best performance and the Object CNN obtains comparable results with the A&C CNN as measured by MRSSE and Pearson's r metric.

Category	Method	MRSSE	Pearson's r
Scene	GIST	0.4756	0.3281
	BOV-SIFT	0.4777	0.3231
	FV-SIFT	0.4570	0.3834
	Scene CNN	0.4084	0.4856
	A&C CNN	**0.3988**	**0.5042**
Object	GIST	0.4956	0.2986
	BOV-SIFT	0.5077	0.2584
	FV-SIFT	0.4843	0.3338
	Object CNN	0.4092	0.4985
	A&C CNN	**0.4025**	**0.5106**
Texture	GIST	0.5843	0.1809
	BOV-SIFT	0.5651	0.2535
	FV-SIFT	0.5448	0.3202
	Texture CNN	0.4567	0.5084
	A&C CNN	**0.4415**	**0.5214**

Table 4: Regression results of aesthetic quality assessment for "scene", "object" and "texture".

4.4. "Texture" results

For evaluation on classification task for "texture", we compare the results of our network with SVM classifiers using GIST, BOV-SIFT and FV-SIFT features, our implemented Rapid SCNN [16], and the architecture of the Scene CNN with the input of local view. The results with different methods for "texture" are also shown in Table 3. We can see that the A&C CNN on the "texture" images performs the best. It is also suggested that the local view is as effective as the global view for the "texture" from the Texture CNN.

For the regression task, the regression network is initialized by the classification model for the Texture CNN. The A&C CNN with regression model is also only tested on the "texture" images. The MRSSE and Pearson's r correlation coefficient for each regression method for "texture" are shown in Table 4. The variance σ^2 of the ground truth in the "texture" test set is 0.6022. We achieve $MRSSE = 0.4567$ with the Texture CNN, and $MRSSE = 0.4415$ with the A&C CNN. Moreover, Table 4 shows that our A&C CNN achieves the best performance and the Texture CNN obtains comparable results with the A&C CNN, as measured by MRSSE and Pearson's r metric.

4.5. Results on overall images

In addition to the evaluation for each specific CNN on each category, we also evaluate our framework on the whole set of images. To demonstrate the effectiveness of our framework for aesthetic classification, we compare our results with that of Rapid SCNN [16], SCNN [16], DCNN [16], RDCN-

Method		Category			
		Scene	**Object**	**Texture**	**Overall images**
$\delta = 1$	[29]	–	–	–	67.00
	Rapid SCNNs [16]	75.44	72.84	64.0	73.29
	SCNN [16]	–	–	–	68.63
	DCNN [16]	–	–	–	73.05
	RDCNN [16]	–	–	–	73.70
	Our framework with three specific CNNs for classification	75.91	73.66	68.8	74.18
	Our A&C CNN for classification	76.04	73.30	71.6	74.13
$\delta = 0$	[29]	–	–	–	66.70
	SCNN [16]	–	–	–	71.20
	DCNN [16]	–	–	–	73.25
	RDCNN [16]	–	–	–	74.46
	Our A&C CNN for classification	76.20	73.91	70.4	74.50
Our framework with three specific CNNs for regression		75.76	73.82	71.1	74.33
Our A&C CNN for regression		76.06	73.89	71.7	**74.51**

Table 5: Accuracy (%) of different methods on overall images.

N [16] and the method in [29]. The results of our framework with three specific CNNs are obtained by combining the results of the three categories using Scene CNN, Object CNN and Texture CNN respectively in Sec. 4.2, 4.3 and 4.4. The results of Rapid SCNNs [16] are obtained by combining the results of our implemented Rapid SCNNs for the three categories. SCNN [16], DCNN [16], RDCNN [16] and the method in [29] are evaluated on the whole set of images without considering their category labels. SCNN [16] is a single-column CNN, DCNN [16] is a double-column CNN with two inputs consisting of a global view and a local view, and RDCNN [16] is a double-column CNN with an aesthetic column and a style column. A&C CNN is evaluated on the whole set of images with considering their category labels. As shown in Table 5, our implemented Rapid SCNN outperforms the method SCNN [16], our framework with three specific CNNs and the A&C CNN outperform the method SCNN [16], DCNN [16] and even RDCNN [16] with adding the photographic style information when $\delta = 1$, which demonstrate the effectiveness of the idea of dividing the images into three categories. For further evaluation on classification model, we compare the A&C CNN with $\delta = 0$ to the state-of-the-art methods in [16]. Our method also obtains the best results, which is comparable to the RDCNN [16]. Moreover, we also evaluate the A&C CNN for category recognition task on the 21000 test images. The accuracies of dividing images into three categories by the A&C CNN with aesthetic classification on $\delta = 1$ and $\delta = 0$ are 83.03% and 85.55% separately, and the accuracy by the A&C CNN with aesthetic regression is 86.20%. It reveals that the A&C CNN performs well on both the tasks of aesthetic quality assessment and category recognition.

To further demonstrate the effectiveness of our framework for regression, the regression results are used for classification. Images in the test set for each category are labeled by thresholding their predicted score with δ set equal to 0, to create the high quality and the low quality classes. As shown in Table 5, classification based on our regression results, especially for the A&C CNN, achieves the comparable performance on the three categories and overall images and outperforms other classification methods on the overall images.

Qualitative results are shown in Fig. 5. The figure contains some test images with ground truths and predicted results using our framework and FV-SIFT for each category. The results predicted by our method are more consistent with the ground truths than the results predicted by FV-SIFT, and our predicted scores contain more information about the degree of aesthetic quality, than can be obtained simply by predicting the class. In contrast to the classification results, the regression results can be very useful in some applications, such as image retrieval. That is, the images can be retrieved based on both their contents and aesthetic scores. The consistency of our classification results and regression results (mapped to classification) is also computed: 91.11% for "scene", 92.01% for "object" and 91.5% for "texture", which suggests that our regression results are compatible with our classification for each category, shown in Fig. 5. Here the consistency is the ratio between the number of the images with same aesthetic class predicted by classification and regression and the total images of each category:

$$consistency = \frac{same\ predictions\ by\ classification\ and\ regression}{total\ images\ of\ each\ category}. \tag{2}$$

From Fig. 5, the consistency of our classification results and regression results for the middle aesthetic images is a little lower, which suggests that it may be more reasonable to label these images with aesthetic scores rather than the class.

To demonstrate the specific features of each category, we apply the trained classification model on one category to the other two categories. For example, the CNN trained on "scene" images is tested on the categories "object" and "texture". Table 6 shows the results, which reveal that the CNNs for "scene" and "object", each trained on its own category, perform best. The CNN for "texture" does not yield the best performance on its own category, probably because its dataset is much smaller than others. However, its accuracy on regression (shown in Table 5) is the best. The filters learned by the first convolutional layer of the three CNNs for aesthetic quality assessments on "scene", "object" and "texture" and the A&C CNN are also shown in Fig. 6. From Fig. 6, although global views are used in two CNNs, the filters of the Scene CNN are much smoother and clearer than those of global column of the Object CNN, which suggests that there are more regions with low frequency information in scene

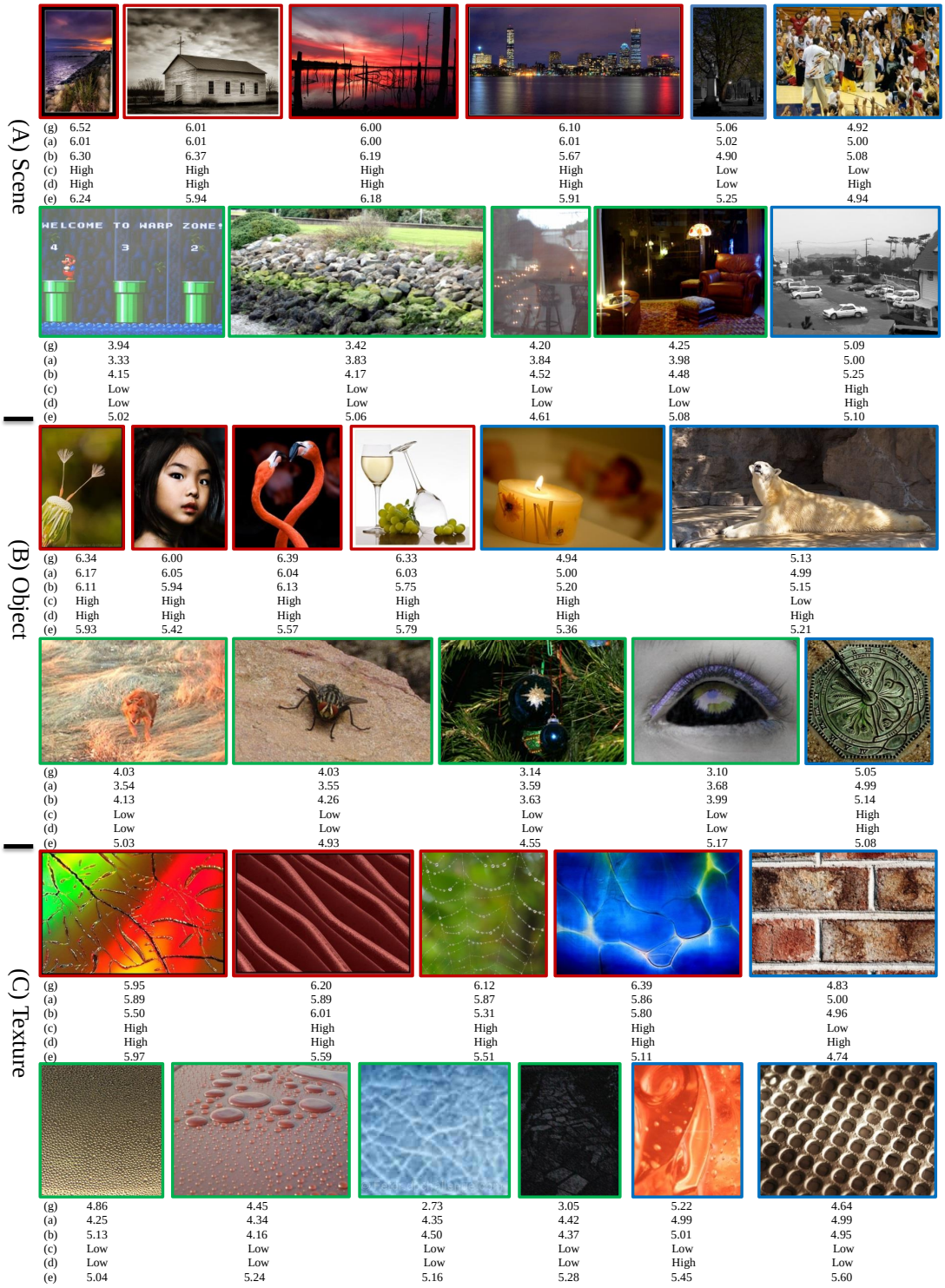
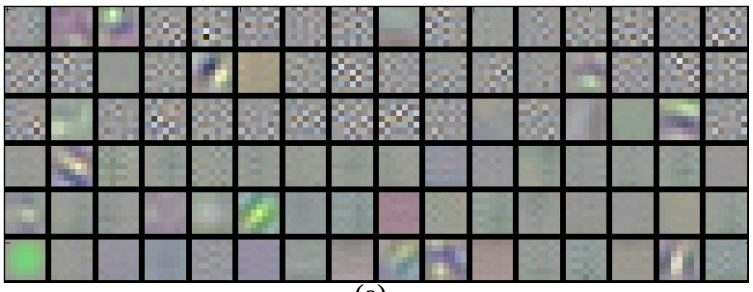
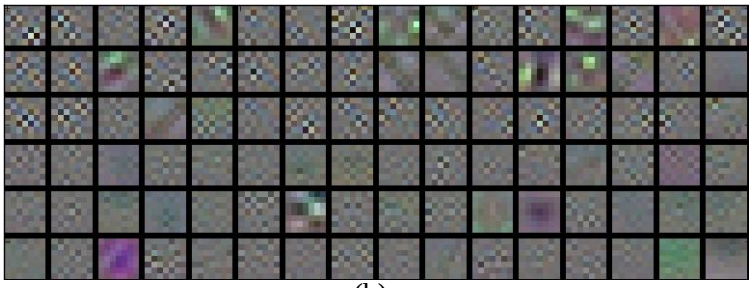


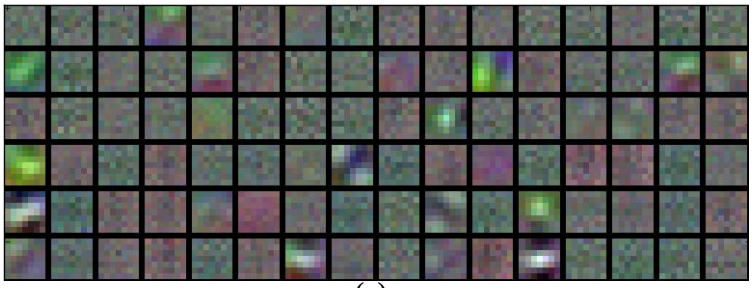
Figure 5: Qualitative results for "scene", "object" and "texture". Each row shows some examples, 4 high (red), 3 medium (blue) and 4 low (green) quality images for each category. (g) The ground truths. (a) The predicted scores of each specific CNN. (b) The predicted scores of the A&C CNN. (c) The predicted class of each specific CNN. (d) The predicted class of the A&C CNN. (e) The predicted scores with FV-SIFT.



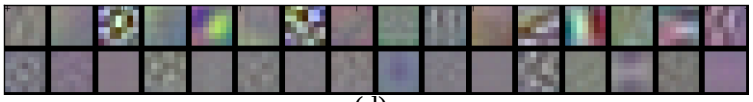
(a)



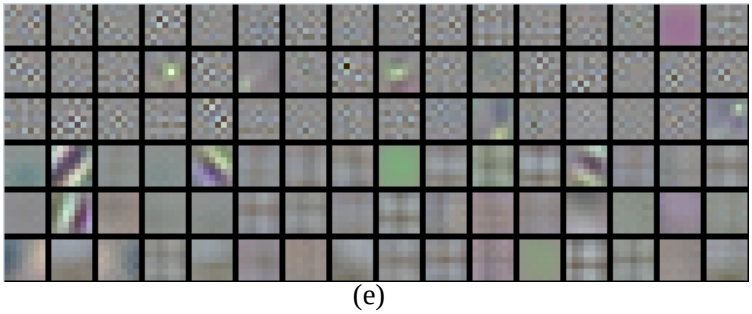
(b)



(c)



(d)



(e)

Figure 6: The filters learned by the first convolutional layer of three CNNs for aesthetic quality assessment on "scene", "object" and "texture". (a) The Scene CNN. (b) The global column of the Object CNN. (c) The salient column of the Object CNN. (d) The Texture CNN. (e) The A&C CNN.

Category	Scene	Object	Texture
Scene model	75.91	70.66	66.9
Single-column Object model	74.14	73.50	69.9
Texture model	73.36	68.51	68.8

Table 6: Accuracy (%) of three categories tested on the classification models with different categories.

than object. For the Object CNN, there are more color changes in the salient column than the global column, which indicates that there is more high frequency information than in the global view. For the texture, there is a great deal of high frequency and color information in the filters of the Texture CNN. In addition, the A&C CNN learns edge and color information for the two tasks. All of these observations suggest that different aesthetic criteria should be used for assessing aesthetic quality in the three categories.

5. Conclusion

In this paper, we propose a novel framework for visual aesthetic quality assessment by dividing the images into three categories: "scene", "object" and "texture". For the three categories, considering their difference on the composition, spatial layout, fixation point and photographic styles etc., three specific CNNs (Scene CNN, Object CNN and Texture CNN) and a simple A&C CNN are designed to learn aesthetic features automatically. In detail, the Scene CNN has an input the global view; the Object CNN has inputs the global view and a salient region for the category "object". In the Texture CNN the local views are the only input. The A&C CNN has inputs the global view and learns features for overall images with considering the both aesthetic and category label. In addition, we interpret aesthetic quality assessment as a classification problem to assess the aesthetic class and also as a regression problem to predict the aesthetic score. We analyze the filters learned by the first convolutional layers in each CNN. Experimental results on the challenging AVA dataset [29] show that aesthetic features learned by the convolutional networks are better than the existing features for aesthetic assessment. Our method outperforms the state-of-the-art methods for each category and for the entire set of images considered without the categories. It is shown that the salient region is very important for assessing the aesthetic quality of "object" images and that the local view is sufficient for assessing "texture" images. In future work, we will investigate those images in each category that have high aesthetic scores.

6. Acknowledgment

This work is funded by the National Basic Research Program of China (Grant No. 2012CB316302), National Natural Science Foundation of China (Grant No. 61322209), the Strategic Priority Research Program of the Chinese Academy of Sciences (Grant No. XDB02050000).

References

[1] D. Joshi, R. Datta, E. Fedorovskaya, Q.-T. Luong, J. Z. Wang, J. Li, J. Luo, Aesthetics and emotions in images, IEEE Signal Processing Magazine 28 (5) (2011) 94–115.

[2] X. Tang, W. Luo, X. Wang, Content-based photo quality assessment, IEEE Transactions on Multimedia 15 (8) (2013) 1930–1943.

[3] L. Marchesotti, N. Murray, F. Perronnin, Discovering beautiful attributes for aesthetic image analysis, International journal of computer vision 113 (3) (2015) 246–266.

[4] R. Datta, J. Li, J. Z. Wang, Learning the consensus on visual quality for next-generation image management, in: Proceedings of the ACM International Conference on Multimedia, 2007, pp. 533–536.

[5] Y. Ke, X. Tang, F. Jing, The design of high-level features for photo quality assessment, in: IEEE International Conference on Computer Vision and Pattern Recognition, Vol. 1, 2006, pp. 419–426.

[6] R. Datta, D. Joshi, J. Li, J. Z. Wang, Studying aesthetics in photographic images using a computational approach, in: European Conference on Computer Vision, 2006, pp. 288–301.

[7] Y. Luo, X. Tang, Photo and video quality evaluation: Focusing on the subject, in: European Conference on Computer Vision, 2008, pp. 386–399.

[8] S. Dhar, V. Ordonez, T. L. Berg, High level describable attributes for predicting aesthetics and interestingness, in: IEEE International Conference on Computer Vision and Pattern Recognition, 2011, pp. 1657–1664.

[9] L. Marchesotti, F. Perronnin, D. Larlus, G. Csurka, Assessing the aesthetic quality of photographs using generic image descriptors, in: IEEE International Conference on Computer Vision, 2011, pp. 1784–1791.

[10] Y. Niu, F. Liu, What makes a professional video? a computational aesthetics approach, IEEE Transactions on Circuits and Systems for Video Technology 22 (7) (2012) 1037–1049.

[11] H.-H. Yeh, C.-Y. Yang, M.-S. Lee, C.-S. Chen, Video aesthetic quality assessment by temporal integration of photo-and motion-based features, IEEE Transactions on Multimedia 15 (8) (2013) 1944–1957.

[12] Y. Wang, Q. Dai, R. Feng, Y.-G. Jiang, Beauty is here: Evaluating aesthetics in videos using multimodal features and free training data, in: Proceedings of ACM international conference on Multimedia, 2013, pp. 369–372.

[13] L. Guo, Y. Xiong, Q. Huang, X. Li, Image esthetic assessment using both hand-crafting and semantic features, Neurocomputing 143 (2014) 14–26.

[14] L. Zhang, Y. Gao, R. Zimmermann, Q. Tian, X. Li, Fusion of multichannel local and global structural cues for photo aesthetics evaluation, IEEE Transactions on Image Processing 23 (3) (2014) 1419–1429.

[15] L. Zhang, Y. Gao, C. Zhang, H. Zhang, Q. Tian, R. Zimmermann, Perception-guided multimodal feature fusion for photo aesthetics assessment, in: Proceedings of the ACM International Conference on Multimedia, 2014, pp. 237–246.

[16] X. Lu, Z. Lin, H. Jin, J. Yang, J. Z. Wang, Rapid: Rating pictorial aesthetics using deep learning, in: Proceedings of the ACM International Conference on Multimedia, 2014, pp. 457–466.

[17] Y. Kao, C. Wang, K. Huang, Visual aesthetic quality assessment with a regression model, in: IEEE International Conference on Image Processing, 2015, pp. 1583 – 1587.

[18] W. Luo, X. Wang, X. Tang, Content-based photo quality assessment, in: IEEE International Conference on Computer Vision, 2011, pp. 2206–2213.

[19] A. Oliva, A. Torralba, Modeling the shape of the scene: A holistic representation of the spatial envelope, International journal of computer vision 42 (3) (2001) 145–175.

[20] A. Oliva, M. L. Mack, M. Shrestha, A. Peeper, Identifying the perceptual dimensions of visual complexity of scenes, in: Proceedings of the 26th Annual Cognitive Science Society, 2004.

[21] O. Axelsson, Towards a psychology of photography: dimensions underlying aesthetic appeal of photographs, Perceptual and Motor Skills 105 (2) (2007) 411–434.

[22] C. E. Nothelfer, K. B. Schloss, S. E. Palmer, The role of spatial composition in preference for color pairs, Journal of Vision 9 (8) (2009) 342–342.

[23] L.-K. Wong, K.-L. Low, Saliency-enhanced image aesthetics class prediction, in: IEEE International Conference on Image Processing, 2009, pp. 997–1000.

[24] L. G. Shapiro, G. C. Stockman, Computer vision, in: Prentice Hall, 2001.

[25] L.-Y. Wei, M. Levoy, Fast texture synthesis using tree-structured vector quantization, in: Proceedings of the 27th annual conference on Computer graphics and interactive techniques, SIGGRAPH '00, 2000, pp. 479–488.

[26] N. Sebe, M. S. Lew, Texture features for content-based retrieval, in: Principles of visual information retrieval, 2001, pp. 51–85.

[27] R. Datta, J. Li, J. Z. Wang, Algorithmic inferencing of aesthetics and emotion in natural images: An exposition, in: IEEE International Conference on Image Processing, 2008, pp. 105–108.

[28] O. Wu, W. Hu, J. Gao, Learning to predict the perceived visual quality of photos, in: IEEE International Conference on Computer Vision, 2011, pp. 225–232.

[29] N. Murray, L. Marchesotti, F. Perronnin, Ava: A large-scale database for aesthetic visual analysis, in: IEEE International Conference on Computer Vision and Pattern Recognition, 2012, pp. 2408–2415.

[30] A. Krizhevsky, I. Sutskever, G. E. Hinton, Imagenet classification with deep convolutional neural networks, in: Advances in neural information processing systems, 2012, pp. 1097–1105.

[31] M. D. Zeiler, R. Fergus, Visualizing and understanding convolutional networks, in: European Conference on Computer Vision, 2014, pp. 818–833.

[32] C. Wang, W. Ren, K. Huang, T. Tan, Weakly supervised object localization with latent category learning, in: European Conference on Computer Vision, 2014, pp. 431–445.

[33] S. Thumfart, R. H. Jacobs, E. Lughofer, C. Eitzinger, F. W. Cornelissen, W. Groissboeck, R. Richter, Modeling human aesthetic perception of visual textures, ACM Transactions on Applied Perception 8 (4) (2011) 27.

[34] K. Huang, Q. Wang, Z. Wu, Color image enhancement and evaluation algorithm based on human visual system, in: IEEE International Conference on Acoustics, Speech, and Signal Processing, Vol. 3, 2004, pp. 721–724.

[35] K. Huang, Z.-y. Wu, G. S. Fung, F. H. Chan, Color image denoising with wavelet thresholding based on human visual system model, Signal Processing: Image Communication 20 (2) (2005) 115–127.

[36] J. Wu, W. Lin, G. Shi, A. Liu, Perceptual quality metric with internal generative mechanism, IEEE Transactions on Image Processing 22 (1) (2013) 43–54.

[37] R. Girshick, J. Donahue, T. Darrell, J. Malik, Rich feature hierarchies for accurate object detection and semantic segmentation, in: IEEE International Conference on Computer Vision and Pattern Recognition, 2014, pp. 580–587.

[38] K. B. Schloss, S. E. Palmer, Aesthetic response to color combinations: preference, harmony, and similarity, Attention, Perception, & Psychophysics 73 (2) (2011) 551–571.

[39] M. Nishiyama, T. Okabe, I. Sato, Y. Sato, Aesthetic quality classification of photographs based on color harmony, in: IEEE International Conference on Computer Vision and Pattern Recognition, 2011, pp. 33–40.

[40] K. Huang, Q. Wang, Z.-Y. Wu, Natural color image enhancement and evaluation algorithm based on human visual system, Computer Vision and Image Understanding 103 (1) (2006) 52–63.

[41] J. Redi, I. Povoa, et al., The role of visual attention in the aesthetic appeal of consumer images: A preliminary study, in: Visual Communications and Image Processing, 2013, pp. 1–6.

[42] M. Nadal, E. Munar, G. Marty, C. Cela-Conde, Visual complexity and beauty appreciation: Explaining the divergence of results, Empirical Studies of the Arts 28 (2) (2010) 173–191.

[43] C. Yang, L. Zhang, H. Lu, X. Ruan, M.-H. Yang, Saliency detection via graph-based manifold ranking, in: IEEE International Conference on Computer Vision and Pattern Recognition, 2013, pp. 3166–3173.

[44] F. Perronnin, C. Dance, Fisher kernels on visual vocabularies for image categorization, in: IEEE International Conference on Computer Vision and Pattern Recognition, 2007, pp. 1–8.

[45] F. Perronnin, J. Sánchez, T. Mensink, Improving the fisher kernel for large-scale image classification, in: European Conference on Computer Vision, 2010, pp. 143–156.

[46] J. Sivic, A. Zisserman, Video google: A text retrieval approach to object matching in videos, in: IEEE International Conference on Computer Vision, 2003, pp. 1470–1477.

[47] G. Csurka, C. Dance, L. Fan, J. Willamowski, C. Bray, Visual categorization with bags of keypoints, in: Workshop on statistical learning in computer vision, ECCV, Vol. 1, 2004, pp. 1–2.

[48] D. G. Lowe, Distinctive image features from scale-invariant keypoints, International journal of computer vision 60 (2) (2004) 91–110.

[49] C.-C. Chang, C.-J. Lin, Libsvm: A library for support vector machines, ACM Transactions on Intelligent Systems and Technology 2 (3) (2011) 27.